Data Management in Multi-center Clinical Trials and the Role of a Nation-wide Computer Network. A 5 Year Evaluation.

M.G.M. Gerritsen, O.E.H. Sartorius, F.M. van Rappard vd Veen, G.T. Meester.
Interuniversity Cardiology Institute of the Netherlands, Utrecht, The Netherlands.

ABSTRACT

Multidisciplinary collaboration in multi-center trials needs a formalized data management structure to ensure true progress monitoring and high quality research data. Cadans, a customized facility for data management, related to the Interuniversity Cardiology Institute of the Netherlands, designed a computer-based data management system for multidisciplinary multi-center collaborative research projects. In this paper we describe the system and the role of integrated access to research databases on a data network. Areas of concern are also discussed.

INTRODUCTION

The Cardiology Data Network Structure (CADANS) provides a framework for the coördination of studies conducted by the Interuniversity Cardiology Institute of the Netherlands (ICIN) and integrates the departments of Cardiology in the eight Dutch universities into a nation-wide research data network. A pilot study carried out 5 years ago defined the scenario for the realisation of this project [1]. The network has a star topology. At Cadans, the center of the star, three IBM RT6150 RISC cpu-based microcomputers (AIX v2.2.1 operating system) are mounted as the central database, mail and file servers for two local area networks of 25 80386 based DOS V5.0 client workstations. Cadans is linked up via Internet Serial Line Protocol to an RT6150 node in each of the eight study centers. This local node serves three 80386 based DOS operated Oracle clients in a local area network. Oracle [2] is the relational database management system, PC/TCP and Idrive [3] are the network software and POPmail [4] is the end user mail application to Internet. SQLforms, a client data entry application, is installed at all research workstations.
This presentation is within the context of a large multi-center trial, the REgression GRowth Evaluation Statin Study. REGRESS is a multi-center study using clinical data, laboratory test results and quantitative coronary angiography (QCA) to establish the effects of Pravastatin [5] on coronary atherosclerosis in men with angina pectoris and serum cholesterol levels between 4 and 8 mmol/l, treated either medically or by coronary angioplasty or bypass surgery. The study is designed as a prospective randomized, triple-blind, placebo-controlled trial in 885 men with baseline coronary cine-angiography and follow-up after 2 years. The QCA is carried out in the central facility coronary lab by the MEDIS Cardiovascular Measurement System [6]. Lesions in at least 9 matching coronary segments from baseline and follow-up angiograms are analysed where possible in 2 projections more then 60 ° apart.

APPROACH

The succesful conduct of such a multi-center clinical study requires a data management system for the entire process of study design, support and progress monitoring [7]. A primary objective of the data management system is to provide high quality data by keeping the number of data errors and missing data as low as possible in all treatment groups, and to assemble the maximum amount of data relevant for final analysis.
To accomplish this goal, a relational database management system [8] was installed at 9 of eleven data collection sites, eight of which are included in the nation-wide Cadans network. Oracle supports a client-server architecture

0195-4210/92/$5.00

divided into two functional units: a front-end or client that provides data entry facilities and a back-end or server that allows clients to access information and maintains the integrity of the database. This client-server architecture allows shared access to each local database and online monitoring of study progress in all participating centers.
The data entry software, written by Cadans in the environment of Oracle's SQLforms, checks for data entry errors and provides selfcoding facilities.
The written protocol of REGRESS was designed in collaboration with all participants [9]. This protocol extensively describes design, methods, purpose, plan of study, trouble shooting and statistical considerations. A manual [10] was also written for the data management system which describes the responsibilities of the study committees, specifies the methods and logistics of data and specimen collection, and sets out guidelines for reporting to the sponsor and handling eventual adverse events in patients.
Much attention was paid to the design of the data collection forms (DCF's) and the 34 data entry screens in order to gain appropriate interpretation and coding of patient data, and most importantly, to make them user-friendly. The DCF's are precoded and they have a simple layout.The contents and layout of the screens are made as identical as possible to that of the DCF's. Data entry itself is kept straightforward by implementing automated coding, data quality validation and navigation throughout screens by software triggers.

ERRORS IN DATA ENTRY

Intermediate analyses monitor the extent of permanently missing data as a relevant parameter for the efficiency of data management. An *at* sign '@' was used for permanently missing data items, a *dollar* sign '$' for a not applicable data item, a *tilde* '~' for not relevant data and a *space* for temporarily missing data. These differences were maintained for reasons of monitoring efficiency during all stages of the study. This type of coding does not interfere with the workstation's data checking facilities. At this moment, August 1993, 885 patients are included in this study with a total of 1.8 million data elements. All baseline and follow-up data of 394 patients have been completely entered in the central database. We found 2.3% of all data to be permanently missing (4505 on 193870 relevant data items). The number of typing errors, defined as real mistakes in data entry during transfer of data from DCF to the workstation, was actually found by editing one out of five completed and checked DCF's (data status 5 or higher) to be 0,23%: 202 errors on 87903 data items in 216 completed DCF's with a mean of 412 data items per DCF.
Consequently, we stand by the decision based on early samples, not to carry out double data entry in the databases, and to perform random checks on one out of twenty forms [11]. If the error level increases to more then 0.25% in a specific sample of completed DCF's, a one out of ten check will be performed.

DATA FLOW

In this integrated data management system one can monitor the primary flow of data and specimens i.e. the transport of data to the central database and the transport of specimens to all laboratories [12].
The information flow dealing with editing and monitoring in all data input sites is also under adequate control. The coordinator at each center enters data in precoded, patient visit related DCF's and sends specimens to the lipid, ultrasound, cine and holtertape laboratories. DCF's are completed by the local research nurse. Two qualified Cadans monitors supplied with "worklists" generated at the Cadans network workstations visit data collection points at all centers to edit the completed DCF's. After acceptance of the reviews, all data are transferred by the local research nurse from the DCF's to a workstation at the data collection point, linked up via the local area network to the the local database server.
After each well defined stage in the process of data entry monitoring, a data status "flag" is automatically set by software and shown at the schedule of events screen as follows: datastatus 1, incomplete data entry; 2, complete data entry; 3, data entry checked 4, ready to send, update impossible; 5, data sent to Cadans; 6, check by editing data of the central lab; 7, check by editing lab data from centers; 8, reserved; 9, screen not available. Data is then backed up to the optical disk, IBM 3363 200mB, mounted in the network and available for intermediate analysis thereby allowing information feedback to the monitors [figure 1].

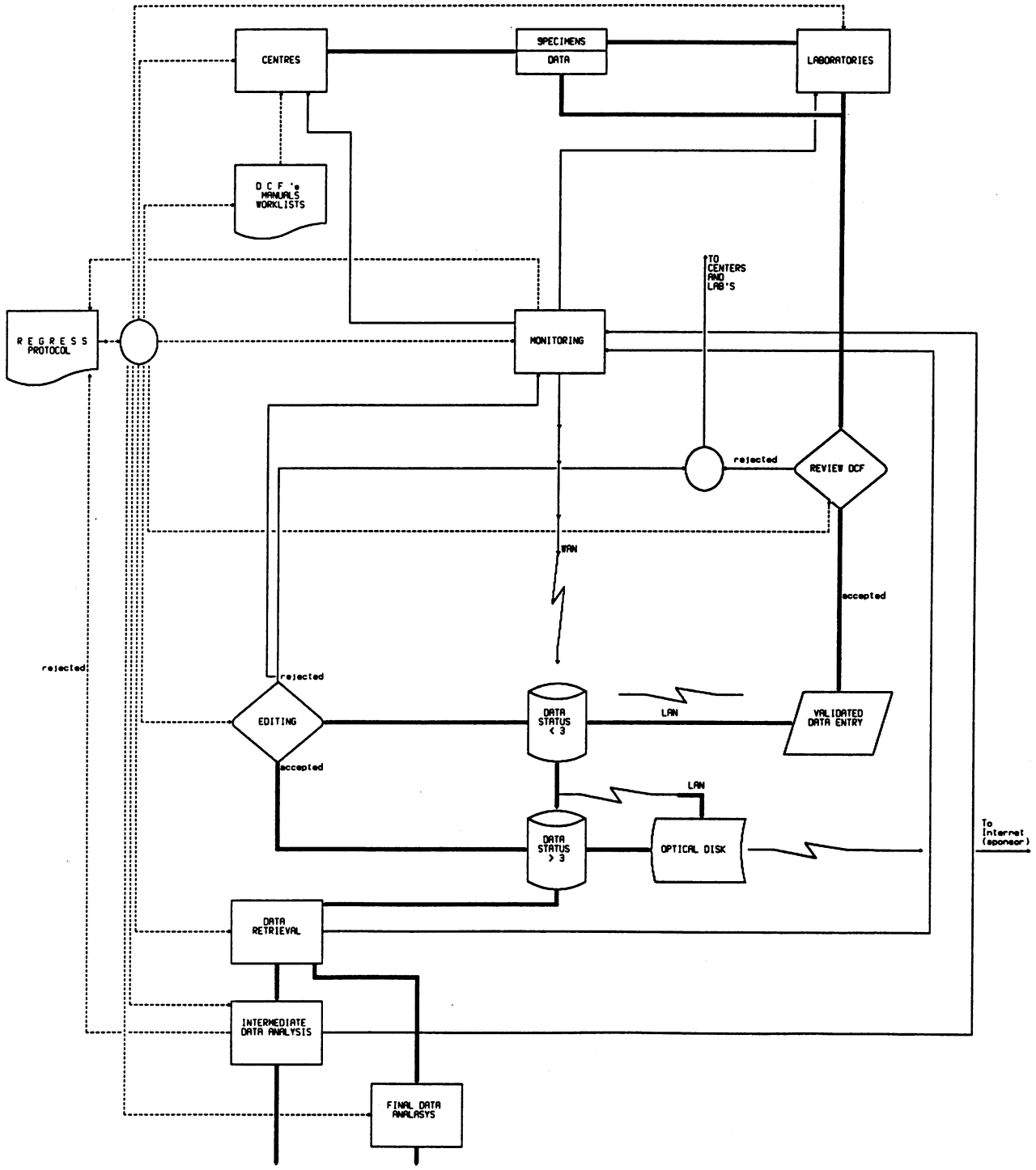

Figure 1. Flow chart of the datamanagement system of Cadans. Bold lines represent primary flow of data and specimens. Solid lines the flow dealing with editing and monitoring. Broken lines the flow of information from and to the protocol. Connection lines represent either WAN or LAN links.

LESSONS LEARNED

Some special areas of concern in conducting multi-center trials should be mentioned:
1) Linking up all centers in a computer network enables local data entry at data collection points and remote data editing in the central facility laboratory, so increasing the potential for correcting errors.
2) A pilot study on the first 20 patients is of course an essential part of a large-scale, collaborative computer-based research project. The first step should be a try out at each data collection point among all data entry personell, to assess error reducing characteristics of the DCF's and screens in practice such as coding features and the extent of user-friendliness. Simplicity and clarity should be the pertinent characteristics of the DCF's and screens.
3) We feel that experienced Intensive Coronary Care nurses are the best monitors for error identification.
4) Ensure that missing data is identified by an efficient detection system and try to keep this system as foolproof as possible.
5) Display related data of previous visits at the workstation during entry of current data, so as to allow checking for possible adverse events in patients, missing data and discontinuity in general.

CONCLUSION

We emphasize the need for extensive computer-based data management to minimize errors and omissions, such as typing errors and missing data, thereby ensuring high quality data. In our view, the secret to successful data management is in the correct design of user-friendly, self coding, clinically oriented forms. The DCF design should incorporate flexibility and entry of free additional text should be possible where coding may lead to misunderstanding or misinterpretation. No serious pitfalls in managing such a complex multi-center trial were detected during the study. We feel that this is due to simultaneous design of protocol, datamanagement lines and database management systems. Moreover, detailed procedure manuals were available to all participants from the outset of the study. There are many benefits of networked workstations linked up to central database servers such as automated data status monitoring, online message facilities, shared access and monitoring, therefore preventing delays, protocol violation and data errors when personnel change in the study period.

References

[1]. M. Gerritsen., G. Meester. Cadans: Implementation of a nation-wide data network for cardiology an initial evaluation. Presented at the seventeenth Annual Symposium on Computers in Cardiology, Chicago, Illinois, September 1990
[2]. Oracle, The Relational Database Management System, version 2.1.32 for workstation and version 5.1.7 for servers, 1989, Oracle Corporation, Belmont, Ca, U.S.A.
[3]. PC/TCP plus, FTP Software Inc., 26 Princess Street, Wakefield, MA
[4]. POPmail/PC, Version 3.1. Copyright 1992 University of Minnesota.
[5]. E.R. Squibb & Sons, Princeton, U.S.A.
[6]. Reiber, J.H.C and P.W. Serruys(eds.) Advances in Quantitative Coronary Arteriography. Kluwer Ac. Publishers, Boston, 1993
[7]. Barnwell, G.M., and Y.N. Marinez. Data managment for collaborative research. Presented at Third Annual Primary Care Research Methods and Statistics Conference, San Antonio, Texas, December 1988
[8]. Meester, Geert T. and F. Pinceroli(eds.) Databases for Cardiology. Kluwer Academic Publishers, Boston, 1991
[9]. REGRESS Protocol #27201-82 September 1990.
[10]. Manual of operations and study organization. Protocol 27201-82, Squibb, September 1990
[11]. Mullooly, John P. The effects of data entry error: an analysis of partial verification. Computers and Biomedical Research 23, 259-267(1990)
[12]. B. Spilker, Guide to Planning and Managing Multiple Clinical Studies. Raven Press, New York, 1987